

### Exercise 1: Multiclass Hinge Loss

Consider the multiclass classification scenario consisting of a feature space  $\mathcal{X}$  and a label space  $\mathcal{Y} = \{1, \dots, g\}$  with  $g \geq 2$  classes. Moreover, we consider the hypothesis space of models based on  $g$  discriminant/scoring functions:

$$\mathcal{H} = \{f = (f_1, \dots, f_g)^\top : \mathcal{X} \rightarrow \mathbb{R}^g \mid f_k : \mathcal{X} \rightarrow \mathbb{R}, \forall k \in \mathcal{Y}\}.$$

A model  $f$  in  $\mathcal{H}$  is used to make a prediction by means of transforming the scores into classes by choosing the class with the maximum score:

$$h(\mathbf{x}) = \arg \max_{k \in \{1, \dots, g\}} f_k(\mathbf{x}). \quad (1)$$

The multiclass hinge loss for models in  $\mathcal{H}$  is defined by

$$L(y, f(\mathbf{x})) = \max_k (f_k(\mathbf{x}) - f_y(\mathbf{x}) + \mathbb{1}_{\{y \neq k\}}).$$

- (a) Show that the 0-1-loss for a predictor  $h$  as in (1) based on a model  $f \in \mathcal{H}$  is at most the multiclass hinge loss for  $f$  i.e.,

$$L_{0-1}(y, h(\mathbf{x})) = \mathbb{1}_{\{y \neq h(\mathbf{x})\}} \leq L(y, f(\mathbf{x})).$$

- (b) Verify that the multiclass hinge loss of  $f \in \mathcal{H}$  on a data point  $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$  is bounded from above by  $\sum_{k \neq y} \max\{0, 1 + f_k(\mathbf{x}) - f_y(\mathbf{x})\}$ .

*Hint:* Note that this upper bound is sometimes referred to as the multiclass hinge loss.

- (c) In the case of binary classification, i.e.,  $g = 2$  and  $\mathcal{Y} = \{-1, +1\}$ , we use a single discriminant model  $f(\mathbf{x}) = f_1(\mathbf{x}) - f_{-1}(\mathbf{x})$  based on two scoring functions  $f_1, f_{-1} : \mathcal{X} \rightarrow \mathbb{R}$  for the prediction by means of  $h(\mathbf{x}) = \text{sgn}(f(\mathbf{x}))$ . Here,  $f_1$  is the score for the positive class and  $f_{-1}$  is the score for the negative class. Show that the upper bound in (b) coincides with the binary hinge loss  $L(y, f(\mathbf{x})) = \max\{0, 1 - yf(\mathbf{x})\}$ .

- (d) Recall the statement of the lecture regarding the binary hinge loss:

“... the hinge loss only equals zero for a margin  $\geq 1$  encouraging confident (correct) predictions.”.

Can we say something similar for the alternative multiclass hinge loss in (b)?

- (e) Now consider the case in which the score functions are linear, i.e.,  $f_k(\mathbf{x}) = \boldsymbol{\theta}_k^\top \mathbf{x}$  for each  $k \in \mathcal{Y}$ . What is the difference between
- a model which is obtained by (empirical) risk minimization of the alternative multiclass hinge loss in (b), and
  - a one-vs-rest model obtained by (empirical) risk minimization of the binary hinge loss for the binary classifiers?